ATLAS Data Acquisition
Jinlong Zhang on behalf of the ATLAS Collaboration

Abstract—ATLAS is a general purpose high energy physics experiment at the Large Hadron Collider. The trigger and data acquisition (TDAQ) system has to handle the extremely high data rates. The three level ATLAS trigger system (level 1, level 2 and event filter) reduces the data rate from 40 MHz bunch crossing down to ∼200 Hz. The DAQ system is designed to transport data across different trigger levels to the mass storage, with main subsystems including the dataflow, a combination of custom designed components and commodity processors running multithreaded software applications and connected by Gigabit Ethernet, the online software responsible for the configuration, control and information sharing of the system, and the monitoring framework responsible for the data quality assurance. The system is being continuously commissioned in situ with detectors, by mainly taking cosmic data and also the beam data in 2008.

I. ATLAS DAQ SYSTEM

The ATLAS detector [1] is designed to study the proton-proton collisions at the CERN Large Hadron Collider (LHC), at a center of mass energy of 14 TeV and bunch crossing rate of 40 MHz. It consists of several highly granular and hermetic concentric subdetector systems, including the magnet system, the inner detector system, the calorimetry system, the muon system and several forward detectors. The magnet configuration consists of a superconducting solenoid and three large superconducting toroids. The inner detector system combines the high resolution semiconductor pixel detector, strip detector and a transition radiation detector. The calorimetry system comprises the high granularity liquid argon electromagnetic sampling calorimeter and the scintillator-iron tile hadronic calorimeter. The muon spectrometer includes high precision tracking chambers and trigger chambers. Forward detectors are mainly used to determine the luminosity.

The ATLAS trigger and DAQ (TDAQ) system has to handle the extremely high data rates (Fig. 1) at the LHC design luminosity. The level 1 (LVL1) trigger reduces the rate down to 75 KHz via the custom built electronics [2]. The level 2 (LVL2) trigger brings the rate further down to ∼3.5 KHz. Finally the event filter (EF) reduces the rate down to ∼200 Hz for permanent storage. Both LVL2 and EF, the High Level Trigger (HLT), run selection algorithms on commodity processors. The DAQ system is designed to transport data to and from the different trigger levels to the mass storage. Its main subsystems are the dataflow and online software, the former being a combination of custom designed components and commodity processors running multithreaded software applications and connected by Gigabit Ethernet. The online software is responsible for the configuration, control, information sharing and the monitoring framework responsible for the data quality assurance.

Manuscript received June 14, 2009.
Jinlong Zhang is with Argonne National Laboratory, Argonne, IL 60439 USA (e-mail: zhangjl@anl.gov).

II. DATAFLOW

Fig. 2 illustrates the baseline of the dataflow architecture [3]. Data fragments of LVL1 accepted events are transferred from the detector front-end readout electronics to the Read Out Systems (ROSs). For each LVL1 accepted event, Regions of Interest (RoI) identified by the LVL1 are collected by the Region of Interest Builder (RoIB), a 9U VMEbus system [5], and provided to the LVL2 supervisors (L2SVs). The latter supervises the handling and processing of an event by the LVL2. LVL2 processing Units (L2PUs) execute trigger selection algorithms on the data obtained from the ROSs defined by the RoI and sends the results of its processing to the LVL2 result handler (L2RH). Data fragments for LVL2 accepted events are then built into single event, under the supervision of the Data Flow Manager (DFM), by the event building applications (SFI) across a switched Gigabit Ethernet network. The SFI’s then send the complete events to the EF nodes (EFD/PT). Events selected by the EF are subsequently sent to the local data storage (SFO) before being transferred to permanent storage for offline reconstruction. Each SFO has a large disk array of 12 TB and uses three file systems via three RAID disk controllers to maximize the throughput.

Most of the element interconnections in the dataflow system are performed with the standard Gigabit Ethernet network and switching technology, an exception being the S-LINK [4] connections between the RoIB and the L2SVs and the detector systems and the ROSs. Each L2SV incorporates a S-LINK to PCI interface card and each ROS contains several Read Out Buffers (ROBINs) [6] each of which incorporates three S-LINK interfaces. The L2SV, L2RH, DFM, SFI, SFO, L2PU and EFD/PT nodes are multi-core PCs, with large CPU power and memory size with respect to individual applications.

As shown in Table I, a large fraction of the dataflow system has been deployed and in particularly the full ROS, SFI and

![Figure 1. Schematic diagram of the TDAQ system.](image-url)
TABLE I
DATAFLOW HARDWARE INSTALLATION STATUS. THE NUMBERS OF NODES FOR EACH SUBSYSTEM ARE SHOWN EXCEPT THE NUMBERS OF CRATES FOR THE RoIB.

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoIB</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L2SV</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>L2PU node</td>
<td>see XPU</td>
<td>500</td>
</tr>
<tr>
<td>DFM</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>SFI</td>
<td>63</td>
<td>63</td>
</tr>
<tr>
<td>EF node</td>
<td>see XPU</td>
<td>1800</td>
</tr>
<tr>
<td>SFO</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>ROS</td>
<td>153</td>
<td>153</td>
</tr>
<tr>
<td>XPU</td>
<td>850</td>
<td>850</td>
</tr>
</tbody>
</table>

SFO farms are available. The currently installed HLT nodes are configured in such a way that the nodes can be used as either LVL2 or EF node (so called XPU nodes). During the installation and commissioning phase the performance of individual components has been evaluated. In Fig. 3, the upper left plot shows the single channel readout rate vs. the input fragment size for the RoIB with different output configurations; the upper right plot shows the request rate and bandwidth vs. the fragment size for the ROS before and after optimization; the lower left plot shows the LVL2 input rate with four L2SVs at the low luminosity and the lower right plot shows the EB/EF rate scaling to the network limit of the installed hardware. Without limitation from the downstream, for a LVL1 fragment size of ~18 words as expected, the RoIB would handle a ~320 KHz LVL1 input rate with 8 output channels. The ROS can handle the 20 KHz request rate as designed, only for the small event fragment size before the system optimization and also for the canonical (~200 words) and larger fragment sizes after the system optimization. Four L2SVs can already handle a ~60 KHz input rate even with a trigger menu for the low luminosity which has much higher LVL2 acceptance than the design spec. Before the full EB capability is explored, the EB throughput rate scales linearly with the number of SFI to the installed EF network bandwidth, ~6 GB/s, which is about half of that for the final EF system. For the canonical event size of 1.5 MB, this is beyond the design EB rate.

III. ONLINE SOFTWARE

The online software provides configurations for TDAQ and detector descriptions, as well coordinates a large number of applications (~25,000 for the final system) in data-taking sessions.

The descriptions of the hardware and software used for data-taking are stored in the configuration database. The database is designed with object-oriented database technology and abstract APIs with plugins available for client applications. To support concurrent access by thousands of applications, and to propagate the run-time changes, remote database servers are deployed to ensure short access times. A graphic user interface (GUI) has been developed for users to modify the database and a database archiving service provides for the browsing and retrieval of run time configurations. Additional services are also implemented to cache query results to other relational databases accessed during data-taking.

Two additional basic components of the online software are the process manager and the run control. The process management daemon on each node manages (launches or interrupts) processes according to requests by other components of the online software, e.g. the run control. In doing so it verifies the validity of the requested action with the access manager and ensures with the resource manager that the required resources are available.

A hierarchical tree of run controllers steers the data-taking session by carrying all elements through a finite state machine and ensures that all components are in a coherent state. The run control GUI in Fig. 4 shows some detector components in running state, and also some run information and messages. The diagnostic and error recovery system is integrated into the run control. The diagnostic system can launch tests to investigate the cause of problems and the recovery system can take corresponding action. The audit service archives all error and information messages in a universal format and tools are available to perform analysis afterward.
IV. MONITORING FRAMEWORK

In order to assess the data quality and detector status, the monitoring framework collects the operational data and performs a analysis. Operational data varies from samples of physics events to detector states, from simple values of parameters to histograms. The collection of these different types of data is performed by the information service (IS), online histogramming service (OHS) and event monitoring service (EMON). IS provides the distribution of the values of simple variables. OHS handles raw and ROOT histograms. EMON provides the functionality to run monitoring or calibration algorithm on partial or complete events in the dataflow system using the offline software framework. The operational data collecting introduces significant network traffic of \(~40\) MB/s at the partial event level and \(~220\) MB/s at the complete event level currently.

To view and analyze the histograms, several presenting tools have been developed. Data quality monitoring framework performs the automatic comparison of newly acquired operational data to reference data (histograms or values), performs statistical comparisons and alarm generation in the case of significant deviations. Trigger presenter provides the time history of rates at different trigger levels and detailed information on specific trigger components can also be checked by navigating through the interface (Fig. 5).

V. COMMISSIONING AND EARLY OPERATION

While the final system is being installed and tested incrementally, functionality tests have been performed on a testbed continuously for many years. These tests have been performed either by the ROSs generating data fragments on request or by the pre-loading of the ROSs with simulated event data or cosmic events. The commissioning focuses on deploying subsystem functionalities, improving subsystem performance, performing stress tests and accumulating operational experience. In the process of continuous support for detector installation and commissioning, invaluable feedback on the functionality and operational aspects of the system has also been obtained.

The full TDAQ chain, including LVL1, RoIB, LVL2, EB and ROS, was tested at the first time in February 2007. A trigger rate of \(30\) Hz at LVL1 for cosmic rays was achieved with partial muon trigger chambers. Downward muons were selected with LVL2 algorithms and accepted events were built and stored with a small event building system. Fig. 6 shows the \(\eta\) and \(\phi\) distribution of the cosmic tracks.

The TDAQ system has been regularly running to take cosmic data, with the detector coverage gradually increasing. During the continuous cosmic run period close to the LHC start-up in 2008, \(~550\) million events have been recorded with the total data size of \(~1\) TB and \(~600,000\) data files. In Fig. 7 the left plot shows partial statistics of that data sample and the right figure is the event display of a cosmic event.

The TDAQ system, together with almost complete detector, is preparing for the LHC beam. With the beam condition changing from cosmic, to single beam, to proton proton collision, different data samples will be used for detectors and TDAQ system to understand and tune the performance. A few iterations will be needed before the optimal strategy can be achieved. The LHC startup luminosity is expected to be \(~10^{31} cm^{-2} s^{-1}\) with less bunches. The initial ATLAS data taking under this condition will focus on commissioning the trigger and detector systems, and studying the basic Standard Model physics signatures. A trigger menu \((10^{31}\) menu\) is being deployed for this running phase, by applying low
thresholds, loose selections and pass-through mode wherever possible. The $10^{31}$ menu has been continuously exercised in the final TDAQ infrastructure with the simulated data. Fig. 8 shows the LVL2 processing time and EF processing time for the accepted events. It is expected that both LVL2 processing time and EF processing time deviate from the design specs because the $10^{31}$ menu is designed to maximize the physics output for the much lower luminosity.

On September 10, 2008 the beam delivery strategy from LHC to ATLAS was to stop the beam on collimators, then to realign with center, then to open collimators and continue. So ATLAS could get one splash event for each beam shot on the collimators. ATLAS relied on the trigger systems in small radius with well defined cosmic timing, i.e., the LVL1 calorimeter trigger and the minimum bias trigger scintillator trigger installed on the calorimeter cryostat. After the initial splash events the beam pickup system as the timing reference was timed in rapidly to trigger on through-going beam. The first splash event seen by the ATLAS detector is shown in Fig. 9.

VI. CONCLUSION

The three-level ATLAS trigger architecture and highly distributed DAQ system have been deployed. A large fraction of dataflow components has been installed and commissioned. The online software has been implemented to meet the challenges. The monitoring framework has been developed and is being widely used. The DAQ system is continuously being used for detector integration and commissioning. The system ran successfully during the first beam period in 2008 and is ready for the incoming beam in the near future.

REFERENCES